Low-rank approximation of linear parabolic equations by space-time tensor Galerkin methods

Thomas Boiveau, Virginie Ehrlacher, Alexandre Ern, Anthony Nouy

Low-rank approximation of linear parabolic equations by space-time tensor Galerkin methods*

Thomas Boiveau† Virginie Ehrlacher† Alexandre Ern†
Anthony Nouy‡

Abstract

We devise a space-time tensor method for the low-rank approximation of linear parabolic evolution equations. The proposed method is a stable Galerkin method, uniformly in the discretization parameters, based on a Minimal Residual formulation of the evolution problem in Hilbert–Bochner spaces. The discrete solution is sought in a linear trial space composed of tensors of discrete functions in space and in time and is characterized as the unique minimizer of a discrete functional where the dual norm of the residual is evaluated in a space semi-discrete test space. The resulting global space-time linear system is solved iteratively by a greedy algorithm. Numerical results are presented to illustrate the performance of the proposed method on test cases including non-selfadjoint and time-dependent differential operators in space. The results are also compared to those obtained using a fully discrete Petrov–Galerkin setting to evaluate the dual residual norm.

Keywords. Parabolic equations, Tensor methods, Proper Generalized Decomposition, Greedy algorithm.

AMS. 65M12, 65M22, 35K20

1 Introduction

The goal of this work is to devise a space-time tensor method for the low-rank approximation of the solution to linear parabolic evolution equations. The method we propose has two salient features. First, it is a stable Galerkin method, uniformly with respect to the discretization parameters in space and in time, leading to quasi-optimal error estimates in the natural norms of the problem as specified below. Second, the method is global in time (and in space), and the approximate solution is iteratively constructed by solving alternatively global problems

*Part of this research was carried out when the first and third authors participated in the Hausdorff Trimester Program "Multiscale Problems: Algorithms, Numerical Analysis and Computation"

†Université Paris-Est, CERMICS (ENPC), 77455 Marne-la-Vallée 2, France and Inria Paris, 75589 Paris, France.

‡GeM - UMR CNRS 6183, École Centrale de Nantes, Université de Nantes, 1 rue de la Noë, 44321 Nantes, France.

in space and in time. More precisely, at iteration $m \in \mathbb{N}^*$, the space-time approximate solution $u^m(x,t)$ is of the form

$$u^m(x,t) = \sum_{1 \leq n \leq m} v^n(x) s^n(t), \tag{1}$$

where v^n and s^n are members of some finite-dimensional spaces V_h and S_k composed of space and time functions having dimension N_h and N_k, respectively. We say that u^m is an approximation of rank m of the exact solution, consisting of a summation of rank-one terms. We employ a greedy rank-one algorithm for computing the sequence of approximations. Specifically, the approximate solution u^m is constructed iteratively, i.e., once u^{m-1} for $m \geq 1$ is known (we set conventionally $u^0 = 0$), the functions v^m and s^m are computed by solving successively a global problem in space and in time having size N_h and N_k, respectively, and which is defined by minimizing some quadratic convex functional. The present method has the potential to be computationally effective if the exact solution can be approximated accurately by low-rank space-time tensors. In this case, space-time compression is meaningful and full parallelism in time can be exploited by the global time solves leading to an overall computational cost of the order of $m(N_h + N_k)$, whereas traditional time-stepping methods are expected to exhibit a cost of the order of $N_h \times N_k$. Thus, computational benefits are expected whenever $m \ll \min(N_h, N_k)$. We notice that in the literature, approximate solutions of the form (1) are typically obtained by a Proper Generalized Decomposition (PGD) [23, 29, 11, 7], which is a greedy algorithm [34]. In the context of parabolic problems, the PGD strategy has been first introduced within the LATIN method in [22]. Theoretical convergence results have been obtained in different contexts, see, e.g., [7, 23, 5, 12, 6]. We leave the question of adaptivity to future work, i.e., the discrete spaces V_h and S_k are fixed a priori in what follows. Possible applications of the present method can be envisaged in optimal control problems constrained by parabolic evolution equations (see, e.g., [17]) and in parabolic evolution equations with random input data (see, e.g., [18]); in both cases indeed, global space-time approaches are important. We also mention the dynamical low-rank integrators from, e.g., [21, 26, 20] which can be used for parabolic evolution equations with the rank referring to the space variables.

Our starting point is the well-posed formulation of the parabolic evolution equation at hand in the setting of space-time Hilbert–Bochner spaces. More precisely, following Lions and Magenes [25, p. 234], the trial space is $X = L^2(I;V) \cap H^1(I;V')$ and the test space is $Z = L^2(I;V) \times L$, where I is the (non-empty, bounded) time interval and the separable real Hilbert spaces (V, L, V') form a Gelfand triple, i.e., $V \hookrightarrow L \equiv L' \hookrightarrow V'$ with densely defined embeddings. The prototypical example is the heat equation for which $V = H_0^1(\Omega)$, $L = L^2(\Omega)$, $V' = H^{-1}(\Omega)$, where Ω is a bounded, Lipschitz, open subset of $\mathbb{R}^d$, $d \geq 1$. More generally, we consider time-dependent linear (possibly non-selfadjoint) operators $A(t) : V \to V'$ that are bounded and coercive, pointwise in time (a.e.). One important assumption for the present method to remain computationally effective is that the space-time operator and source term in the parabolic evolution equation admit a separated representation in space and in time with a relatively low rank, see Eq. (7) below. This assumption is met in practice for a wide range of problems coming from the engineering and

applied sciences. In several situations, it is also possible to devise such low-rank separated representations with good accuracy using the Empirical Interpolation Method (EIM) [4]. The above well-posed space-time formulation allows one to view the parabolic evolution problem as a Minimal Residual (MinRes) formulation, where the exact solution is the unique minimizer over the trial space X of the (square of the) dual norm of the residual in Z'. The present approximation method is formulated as a Galerkin method for the MinRes formulation. More precisely, we look for the minimizer over a finite-dimensional subspace $X_{hk} \subset X$ of the (square of the) dual norm of the residual measured with respect to a space semi-discrete subspace $Z_h \subset Z$. Here, X_{hk} is a linear space generated using space-time tensor-products of elements of the finite-dimensional subspaces $V_h \subset V$ and $S_k \subset H^1(I)$ considered above, whereas $Z_h = (V_h \otimes L^2(I)) \times V_h$, i.e., the time variable is not discretized in Z_h.

Let us put our work in perspective with the literature. The subject of numerical methods to approximate parabolic evolution equations is extremely rich. One important class of methods are the traditional time-stepping schemes which approximate the solution on a succession of time sub-intervals by marching along the positive time direction, see, e.g., [35] for an overview. In the context of parabolic evolution equations, implicit schemes are often preferred to circumvent the classical CFL restriction on explicit schemes, which is of the form $\delta_k \lesssim \delta_h^2$ (where $\delta_k \sim N_k^{-1}$ is the time-step and $\delta_h \sim N_h^{-1/d}$ is the space-step), but at the expense of having to solve a large linear system of equations at each time-step. The cost of having to solve these systems sequentially has motivated the devising of parareal methods [24, 13] based on iterative corrections at all time sub-intervals simultaneously using the global space-time discrete system associated with time-stepping methods. This global space-time discrete system has also been central in the devising of space-time domain decomposition methods based on waveform relaxation [19, 14, 15]. In contrast to the above approaches which do not make a direct use of the well-posed functional setting in space-time Hilbert–Bochner spaces, a space-time adaptive wavelet method for parabolic evolution problems was proposed and analyzed in [32], involving a rather elaborate construction of the wavelet bases. Simpler hierarchical wavelet-type tensor bases on a space-time sparse grid were also considered in [16] within a heuristic space-time adaptive algorithm, but without offering guaranteed a-priori stability, uniformly with respect to the discretization parameters. We also mention the recent work [28] where the above functional setting for parabolic evolution problems is used to devise preconditioned time-parallel iterative solvers. Furthermore, PGD approximations based on a discrete MinRes formulation measured in the space-time Euclidean norm of the components in a basis of X_{hk} have been devised and evaluated numerically in [30], obtaining promising results on various model parabolic evolution problems.

More recently, space-time Petrov–Galerkin discretizations of parabolic evolution equations were proposed and analyzed in the PhD Thesis [1] and in the related papers [2, 3]. Therein, the same MinRes formulation is considered as in the present work at the continuous level, and the approximate solution is typically sought in the same space-time discrete space. There are, however, several salient differences between [2, 3] and the present work. First, in [2, 3], the dual residual norm of the discrete solution is measured with respect to a *fully discrete* space-time test space, leading to a Petrov–Galerkin formulation,

whereas we consider a *space semi-discrete* test space, leading to a standard Galerkin formulation. The difference is that a careful design of the test space is necessary in the Petrov–Galerkin setting. Precise results in this direction have been obtained in [2]. Let us for instance notice that, for a constant-in-time and selfadjoint differential operator in space (e.g., for the heat equation), the Crank–Nicolson scheme (obtained using continuous piecewise affine time-functions in the trial space and piecewise constant time-functions in the test space) is only conditionally stable, with a stability constant degenerating with the parabolic CFL condition, whereas unconditional stability is achieved by further refining the time mesh used to build the discrete test space as shown in [1, Sec. 5.2.3]. In contrast with this, the present formulation automatically inherits uniform stability with respect to the time-step size. In particular, Lemma 3.1 below shows well-posedness and Lemma 3.3 a quasi-optimal error bound with a constant independent on the time discretization. Nonetheless, the condition number of the discrete matrices should behave as $O(N_k^{-2})$, so that, as usual, the precision can be limited in practice by round-off errors. The second difference lies in the way the discrete system of linear equations is solved iteratively: we use a greedy algorithm to build a sequence of approximate solutions having the form (1), whereas (a generalization of) the LSQR algorithm [31] is used in [3]. Third, we report numerical results on a larger set of model parabolic evolution problems, including in particular non-selfadjoint operators of advection-diffusion-type at moderate Péclet numbers. Finally, let us mention, as already observed in [1, 2, 3], that the inner product with which we equip the discrete test space Z_h plays the role of a preconditioner in the discrete system of linear equations resulting from the discrete MinRes formulation. Equipping Z_h with the natural norm leads to the appearance of the inverse of the stiffness matrix in space. Herein, we explore numerically the effect of relaxing the use of this preconditioner by simply equipping the space-part of Z_h with the Euclidean norm of the components in a basis of V_h. This corresponds to the approach considered in [30].

In Section 2, we specify the functional setting for parabolic evolution equations and the MinRes formulation. This formulation is the basis for the discrete MinRes Galerkin formulation devised in Section 3, where one key idea is the use of a space semi-discrete test space to measure the dual norm of the residual. In Section 4, we present the greedy algorithm we consider to obtain a low-rank approximation of the discrete solution. In Section 5, we present two other discrete MinRes formulations, one using a fully discrete Petrov–Galerkin setting as in [1, 2, 3] and one using the same space semi-discrete setting as in Section 3 but equipping the test space with the above-mentioned Euclidean norm. These additional formulations are introduced for the purpose of performing numerical comparisons. Numerical results on various test cases are discussed in Section 6. Finally, conclusions are drawn in Section 7.

2 Minimal Residual formulation of parabolic evolution equations

In this section, we present the MinRes formulation of parabolic equations, which is at the heart of the tensor approximation method we propose later on.

2.1 Parabolic equations

The functional setting for parabolic equations is well understood (see, e.g., the textbooks by Lions and Magenes [25, p. 234], Dautray and Lions [8, p. 513], and Wloka [38, p. 376]). Let

$$V \hookrightarrow L \equiv L' \hookrightarrow V' \tag{2}$$

be a Gelfand triple where V and L are separable real Hilbert spaces respectively equipped with inner products $\langle\cdot,\cdot\rangle_V$ and $\langle\cdot,\cdot\rangle_L$, with associated norms $\|\cdot\|_V$ and $\|\cdot\|_L$. The symbol $\hookrightarrow$ represents a densely defined and continuous embedding. Let $T > 0$ be the time horizon and let $I := (0, T)$ be the time interval. Let $A : I \to \mathcal{L}(V, V')$ be a strongly measurable time-function with values in the Hilbert space of bounded linear operators from V to V'. We assume that the following boundedness and coercivity properties hold true: there exist $0 < \alpha \leq M < +\infty$ such that a.e. $t \in I$,

$$\|A(t)v\|_{V'} \leq M\|v\|_V, \qquad \forall v \in V, \tag{3a}$$

$$\langle A(t)v, v\rangle_{V',V} \geq \alpha\|v\|_V^2, \qquad \forall v \in V. \tag{3b}$$

We do not require $A(t)$ to be selfadjoint.

Let us define the Hilbert–Bochner spaces

$$X := L^2(I; V) \cap H^1(I; V'), \qquad Y := L^2(I; V), \qquad Z := Y \times L. \tag{4}$$

Since $X \hookrightarrow \mathcal{C}^0(\overline{I}; L)$ with $\overline{I} = [0, T]$, the value at any time $t \in \overline{I}$ of any function $x \in X$ is well-defined as an element of L. In particular, we denote $x(0) \in L$ the initial value of x at the time $t = 0$. The spaces X and Z are equipped with the norms

$$\|x\|_X^2 := \|x\|_{L^2(I;V)}^2 + M^{-2}\|\partial_t x\|_{L^2(I;V')}^2 + \alpha^{-1}\|x(T)\|_L^2, \qquad \forall x \in X, \tag{5a}$$

$$\|z\|_Z^2 := \|y\|_{L^2(I;V)}^2 + \alpha^{-1}\|g\|_L^2, \qquad \forall z = (y, g) \in Y \times L = Z, \tag{5b}$$

where the various scaling factors are introduced to be dimensionally consistent.

Let $f \in Y' = L^2(I; V')$ and let $u_0 \in L$. We consider the following parabolic problem: find $u \in X$ such that

$$\begin{cases} \partial_t u(t) + A(t)u(t) = f(t), & \text{in } V', \text{ a.e. } t \in I, \\ u(0) = u_0, & \text{in } L. \end{cases} \tag{6}$$

For the present space-time tensor methods to be computationally effective, we assume that the operator A and the source term f have the following separated form:

$$A(t) = \sum_{1 \leq p \leq P} \mu^{(p)}(t) A^{(p)} \in \mathcal{L}(V, V'), \qquad f(t) = \sum_{1 \leq q \leq Q} \lambda^{(q)}(t) f^{(q)} \in V', \tag{7}$$

for some positive integers P, Q taking moderate values, where $\mu^{(p)} \in L^\infty(I)$, $A^{(p)} \in \mathcal{L}(V, V')$ for all $1 \leq p \leq P$, and $\lambda^{(q)} \in L^2(I)$, $f^{(q)} \in V'$ for all $1 \leq q \leq Q$. A similar decomposition is considered, e.g., in [27] for the space-time isogeometric discretization of parabolic problems with varying coefficients.

Example 2.1 (Heat equation)**.** *Let Ω be a Lipschitz domain in $\mathbb{R}^d$, $V = H^1_0(\Omega)$, $L = L^2(\Omega)$, and $V' = H^{-1}(\Omega)$. Let $\mu : I \to \mathbb{R}$ be a measurable function bounded from above and from below away from zero uniformly on I. Then, the time-dependent family of operators $(A(t))_{t\in I}$ defined such that $A(t) := -\mu(t)\Delta$, for a.e. $t \in I$, where $\Delta \in \mathcal{L}(V; V')$ is the Laplacian operator, satisfies the assumptions* (3) *and* (7)*. Whenever $\mu(t) \equiv 1$, the family of operators is time-independent, and one recovers the prototypical example of the heat equation.*

2.2 Well-posedness and minimal residual formulation

It is convenient to introduce the operator $\mathcal{A} : X \to Y'$ so that

$$(\mathcal{A}x)(t) = \partial_t x(t) + A(t)x(t) \in V', \qquad \text{a.e. } t \in I. \tag{8}$$

Problem (6) can be equivalently rewritten using the operator $\mathcal{B} : X \to Z' = Y' \times L$ such that

$$\mathcal{B}x = (\mathcal{A}x, x(0)) \in Z', \qquad \forall x \in X. \tag{9}$$

Then, an equivalent reformulation of problem (6) reads as follows: find $u \in X$ such that

$$\mathcal{B}u = (f, u_0) \quad \text{in } Z'. \tag{10}$$

It is well-known that the operator $\mathcal{B}$ is bounded, i.e., $\mathcal{B} \in \mathcal{L}(X, Z')$, and satisfies the following two properties:

$$\exists \beta > 0 \quad \text{s.t.} \quad \inf_{x\in X} \sup_{z\in Z} \frac{\langle \mathcal{B}x, z\rangle_{Z',Z}}{\|x\|_X \|z\|_Z} \geq \beta, \tag{11a}$$

$$\forall z \in Z, \quad (\langle \mathcal{B}x, z\rangle_{Z',Z} = 0,\ \forall x \in X) \implies (z = 0), \tag{11b}$$

where it is implicitly understood that nonzero arguments are considered in the inf-sup condition, and where, for all $(y, g) \in Y \times L = Z$,

$$\begin{aligned}\langle \mathcal{B}x, (y, g)\rangle_{Z',Z} &= \langle \mathcal{A}x, y\rangle_{Y',Y} + \langle x(0), g\rangle_L \\ &= \int_I \langle \partial_t x(t) + A(t)x(t), y(t)\rangle_{V',V} dt + \langle x(0), g\rangle_L. \end{aligned} \tag{12}$$

Therefore, owing to the Banach–Nečas–Babuška Theorem (see, e.g., [9, Thm. 2.6]), $\mathcal{B}$ is an isomorphism. The proof of the well-posedness of parabolic problems by means of inf-sup arguments can be found in [9, Thm. 6.6] using a strongly enforced initial condition; a systematic treatment can be found more recently in [33]. Since the operator norm of $\mathcal{B}$ and the inf-sup constant β in (11a) play an important role in what follows, we provide respectively an upper bound and a lower bound for these two constants. For a.e. $t \in I$, the inverse adjoint operator $A(t)^{-\mathrm{T}}$ is well-defined in $\mathcal{L}(V'; V)$, and we have $M^{-1}\|\varphi\|_{V'} \leq \|A(t)^{-\mathrm{T}}\varphi\|_V \leq \alpha^{-1}\|\varphi\|_{V'}$ and $\langle \varphi, A(t)^{-\mathrm{T}}\varphi\rangle_{V',V} \geq \frac{\alpha}{M^2}\|\varphi\|^2_{V'}$, for all $\varphi \in V'$.

Lemma 2.2 (Boundedness)**.** *The norm of the operator $\mathcal{B} : X \to Z'$ is such that*

$$\|\mathcal{B}\|_{\mathcal{L}(X;Z')} := \sup_{x\in X} \sup_{z\in Z} \frac{\langle \mathcal{B}x, z\rangle_{Z',Z}}{\|x\|_X \|z\|_Z} \leq \sqrt{3}M.$$

Proof. For all $x \in X$, $z = (y, g) \in Z = Y \times L$, we have

$$\begin{aligned}\langle \mathcal{B}x, z\rangle_{Z',Z} &= \langle \mathcal{A}x, y\rangle_{Y',Y} + \langle x(0), g\rangle_L \le \|\mathcal{A}x\|_{Y'}\|y\|_Y + \|x(0)\|_L\|g\|_L \\ &\le \left(\|\mathcal{A}x\|_{Y'}^2 + \alpha\|x(0)\|_L^2\right)^{1/2}\left(\|y\|_Y^2 + \alpha^{-1}\|g\|_L^2\right)^{1/2} \le \sqrt{3}M\|x\|_X\|z\|_Z,\end{aligned}$$

since $\|\mathcal{A}x\|_{Y'}^2 \le 2M^2(M^{-2}\|\partial_t x\|_{Y'}^2 + \|x\|_Y^2) \le 2M^2\|x\|_X^2$ and $\alpha\|x(0)\|_L^2 \le \alpha\|x(T)\|_L^2 + 2\alpha\|\partial_t x\|_{Y'}\|x\|_Y \le \alpha\|x(T)\|_L^2 + \alpha M(M^{-2}\|\partial_t x\|_{Y'}^2 + \|x\|_Y^2) \le M^2\|x\|_X^2$ (recall that $\alpha \le M$). □

Lemma 2.3 (Inf-sup constant)**.** (11a) *holds true with the inf-sup constant* $\beta \ge \frac{\alpha^2}{M}(1+\kappa^2)^{-\frac{1}{2}}$*, where* $\kappa := \operatorname{ess\,sup}_{t\in I}\frac{1}{2}\|A(t)A(t)^{-\mathrm{T}} - I\|_{\mathcal{L}(V';V')}$.

Proof. For completeness, we briefly outline the proof which is similar to that of [33, Prop. 3.1], but with a different scaling for the norms. Let $x \in X$ and let us take

$$z = (A(t)^{-\mathrm{T}}\partial_t x(t) + x(t), x(0)) \in Z.$$

Then, using the coercivity of $A(t)$ and of $A(t)^{-1}$, we have

$$\begin{aligned}\langle \mathcal{B}x, z\rangle_{Z',Z} &= \int_I \langle \partial_t x(t) + A(t)x(t), A(t)^{-\mathrm{T}}\partial_t x(t) + x(t)\rangle_{V',V}dt + \|x(0)\|_L^2 \\ &\ge \alpha M^{-2}\|\partial_t x\|_{L^2(I;V')}^2 + \alpha\|x\|_{L^2(I;V)}^2 + \|x(T)\|_L^2 \ge \alpha\|x\|_X^2.\end{aligned}$$

Moreover, using again the coercivity of $A(t)$ and the boundedness of $A(t)$ and $A(t)^{-\mathrm{T}}$, we have

$$\begin{aligned}\|z\|_Z^2 &= \int_I \|A(t)^{-\mathrm{T}}\partial_t x(t) + x(t)\|_V^2 dt + \alpha^{-1}\|x(0)\|_L^2 \\ &\le \alpha^{-1}\int_I \langle A(t)\big(A(t)^{-\mathrm{T}}\partial_t x(t) + x(t)\big), A(t)^{-\mathrm{T}}\partial_t x(t) + x(t)\rangle_{V',V}dt + \alpha^{-1}\|x(0)\|_L^2 \\ &\le M\alpha^{-1}\|x\|_{L^2(I;V)}^2 + M^2\alpha^{-2}\|\partial_t x\|_{L^2(I;V')}^2 + \alpha^{-1}\|x(T)\|_L^2 \\ &\quad + 2\kappa\alpha^{-1}\|\partial_t x\|_{L^2(I;V')}\|x\|_{L^2(I;V)} \\ &\le (1+\kappa^2)M^2\alpha^{-2}\|x\|_X^2,\end{aligned}$$

and the conclusion is straightforward. □

Remark 2.4 (Heat equation)**.** *Sharp estimates of the inf-sup constant* β *for the heat equation (with* $\mu(t) \equiv 1$ *on* I *so that* $\alpha = M = 1$ *and* $\kappa = 0$*) can be found in [36, 10] using the above norms.*

The solution to the parabolic equation (6) is the global minimizer of the residual-based quadratic functional $\mathcal{E} : X \to \mathbb{R}$, defined such that

$$\mathcal{E}(x) := \frac{1}{2}\|\mathcal{B}x - (f, u_0)\|_{Z'}^2, \qquad \forall x \in X, \tag{13}$$

where we equip the space $Z' = Y' \times L$ with the norm

$$\|(\phi, g)\|_{Z'}^2 := \|\phi\|_{L^2(I;V')}^2 + \alpha\|g\|_L^2. \tag{14}$$

Since the functional $\mathcal{E}$ is strongly convex on X with parameter $\beta^2 > 0$ owing to the inf-sup condition (11a), $\mathcal{E}$ admits a unique global minimizer in X, and since the operator $\mathcal{B}$ is surjective, the minimum value of $\mathcal{E}$ on X is zero. In other words, the unique solution to (6) can be equivalently characterized as follows:

$$u = \operatorname*{argmin}_{x\in X}\frac{1}{2}\left(\|\mathcal{A}x - f\|_{Y'}^2 + \alpha\|x(0) - u_0\|_L^2\right). \tag{15}$$

3 Discrete Minimal Residual Galerkin formulation

In this section, we introduce a discrete energy based on a space semi-discrete Galerkin method to approximate the unique minimizer in (15). We consider finite-dimensional spaces V_h and S_k such that

$$V_h \subset V, \qquad S_k \subset S := H^1(I), \tag{16}$$

and we set $N_h := \dim(V_h)$ and $N_k := \dim(S_k)$. Typically, V_h is constructed using $\mathbb{P}_1$ Lagrange finite elements on a space mesh of Ω and S_k is constructed using continuous, piecewise affine functions on a time mesh of I. We are going to seek the discrete minimizer in the tensor-product space

$$X_{hk} := V_h \otimes S_k \subset X, \tag{17}$$

which is of dimension $\dim(X_{hk}) = N_h \times N_k$.

3.1 Space semi-discrete Galerkin approximation

Let us set

$$X_h = V_h \otimes H^1(I), \quad Y_h := V_h \otimes L^2(I) \equiv L^2(I; V_h), \quad Y_h' = V_h' \otimes L^2(I) \equiv L^2(I; V_h'). \tag{18}$$

Since V_h is a subspace of V and owing to (2), we have $Y_h \subset Y$ and $Y' \subset Y_h'$, where the second inclusion follows by restricting the action of linear forms on V to V_h. Let us define $f_h \in Y_h'$ s.t. $f_h(t) = f(t)|_{V_h}$ a.e. $t \in I$. Recalling the separated form (7), we have

$$f_h(t) := \sum_{1 \le q \le Q} \lambda^{(q)}(t) f^{(q)}|_{V_h} \qquad \text{a.e. } t \in I. \tag{19}$$

Let u_{0h} be the L-orthogonal projection of u_0 onto V_h. Let $\mathcal{A}_h : X_h \to Y_h'$ be s.t.

$$(\mathcal{A}_h x_h)(t) := (J_{V_h} \otimes \partial_t) x_h(t) + A_h(t) x_h(t) \in V_h', \qquad \text{a.e. } t \in I, \tag{20}$$

where $J_{V_h} : V_h \to V_h'$ is the injection resulting from (2), i.e., $\langle J_{V_h} v_h, w_h \rangle_{V_h', V_h} = \langle v_h, w_h \rangle_L$ for all $v_h, w_h \in V_h$, and $A_h(t) : V_h \to V_h'$ is the discrete counterpart of $A(t)$ s.t. $\langle A_h(t) v_h, w_h \rangle_{V_h', V_h} = \langle A(t) v_h, w_h \rangle_{V', V}$. Let us introduce the space semi-discrete space $Z_h = Y_h \times L_h$ so that $Z_h' = Y_h' \times L_h'$, where L_h coincides with V_h as linear space but is equipped with the norm of L (note that $L_h \subset L \equiv L' \subset L_h'$). Note that $Z_h \subset Z$. Let $\mathcal{B}_h : X_h \to Z_h'$ be the operator defined for $x_h \in X_h$ by

$$\mathcal{B}_h x_h = (\mathcal{A}_h x_h, x_h(0)) \in Z_h' = Y_h' \times L_h', \tag{21}$$

and such that, for all $(y_h, g_h) \in Y_h \times L_h$, (compare with (12))

$$\langle \mathcal{B}_h x_h, (y_h, g_h) \rangle_{Z_h', Z_h} = \langle \mathcal{A}_h x_h, y_h \rangle_{Y_h', Y_h} + \langle x_h(0), g_h \rangle_L = \langle \mathcal{B} x_h, (y_h, g_h) \rangle_{Z', Z}. \tag{22}$$

The space semi-discrete formulation is as follows: find $u_h \in X_h$ such that

$$\mathcal{B}_h u_h = (f_h, u_{0h}) \quad \text{in } Z_h'. \tag{23}$$

The well-posedness of this formulation is ensured by the following lemma, where the subspace $X_h \subset X$ is equipped with the norm $\|x_h\|_{X_h}^2 := \|x_h\|_{L^2(I;V)}^2 + M^{-2}\|\partial_t x_h\|_{L^2(I;V_h')}^2 + \alpha^{-1}\|x_h(T)\|_L^2$, and the subspace $Z_h \subset Z$ is equipped with the norm of Z. Recall the inf-sup constant $\beta = \frac{\alpha^2}{M}(1+\kappa^2)^{-\frac{1}{2}}$ from Lemma 2.3.

Lemma 3.1. *The operator $\mathcal{B}_h : X_h \to Z_h'$ satisfies $\|\mathcal{B}_h\|_{\mathcal{L}(X_h;Z_h')} \le \sqrt{3}M$, and*

$$\inf_{x_h \in X_h} \sup_{z_h \in Z_h} \frac{\langle \mathcal{B}_h x_h, z_h\rangle_{Z_h',Z_h}}{\|x_h\|_{X_h}\|z_h\|_Z} \ge \beta, \tag{24a}$$

$$\forall z_h \in Z_h, \quad (\langle \mathcal{B}_h x_h, z_h\rangle_{Z_h',Z_h} = 0,\ \forall x_h \in X_h) \implies (z_h = 0). \tag{24b}$$

Proof. The upper bound on $\|\mathcal{B}_h\|_{\mathcal{L}(X_h;Z_h')}$ is shown as in the proof of Lemma 2.2. To prove (24a), one can use the same arguments as in the proof of Lemma 2.3 by picking in the supremizing set $z_h = (A_h(t)^{-\mathrm{T}}\partial_t x_h(t) + x_h(t), x_h(0)) \in Y_h \times L_h = Z_h$. Finally, one can prove (24b) using the following arguments, as in [9, Thm. 6.6]. Let $z_h = (y_h, g_h) \in Z_h$. Taking x_h arbitrary in $V_h \otimes C_0^\infty(I)$ shows that $(J_{V_h} \otimes \partial_t)y_h - A_h^{-\mathrm{T}} y_h = 0$ in Y_h'. Taking next $x_h = tv_h$ with v_h arbitrary in V_h proves that $y_h(T) = 0$, and taking $x_h = ty_h$, one concludes that $y_h = 0$. Finally, taking $x_h = v_h$ with v_h arbitrary in V_h yields $g_h = 0$. □

Lemma 3.1 implies that $\mathcal{B}_h : X_h \to Z_h'$ is an isomorphism such that

$$\beta\|x_h\|_{X_h} \le \|\mathcal{B}_h x_h\|_{Z_h'} \le \sqrt{3}M\|x_h\|_{X_h}, \tag{25}$$

for all $x_h \in X_h$. The solution u_h to the equation (23) is the unique minimizer of the discrete energy functional $\mathcal{E}_h : X_h \to \mathbb{R}$ defined for all $x_h \in X_h$ by

$$\mathcal{E}_h(x_h) := \frac{1}{2}\|\mathcal{B}_h x_h - (f_h, u_{0h})\|_{Z_h'}^2 = \frac{1}{2}\left(\|\mathcal{A}_h x_h - f_h\|_{Y_h'}^2 + \alpha\|x_h(0) - u_{0h}\|_L^2\right). \tag{26}$$

An important property of this discrete energy functional is the strong convexity that is inherited from the continuous setting, uniformly with respect to the space discretization parameter. More precisely, the functional $\mathcal{E}_h$ is strongly convex on X_h with parameter $\beta^2 > 0$ owing to the inf-sup condition (24a). Since the operator $\mathcal{B}_h$ is surjective, the minimum value of $\mathcal{E}_h$ on X_h is zero and is attained at u_h.

Remark 3.2 (Norm $\|\cdot\|_{X_h}$)**.** *The difference between the $\|\cdot\|_X$-norm and the $\|\cdot\|_{X_h}$-norm lies in the use of the dual norm in V_h' and not in V' to measure the time-derivative. Note that $\|x_h\|_{X_h} \le \|x_h\|_X$, for all $x_h \in X_h$. The reason for this difference is that, as shown in [33], the equivalence of the two norms, uniformly with respect to the space discretization, holds true if and only if the L-orthogonal projection onto V_h is V-stable. This uniform stability (with $V = H_0^1(\Omega)$ and $L = L^2(\Omega)$) is, in turn, not known to hold true if general shape-regular meshes are used to build the finite element space V_h; it does hold true if quasi-uniform meshes are used (as it is the case in the present numerical experiments). We emphasize that the use of a discrete dual norm to measure the time-derivative is a general feature that arises in the quasi-optimality of space semi-discrete Galerkin methods for parabolic evolution problems [33], and is not specific to the present setting.*

3.2 Minimal residual Galerkin approximation

An approximation $u_{hk} \in X_{hk}$ of $u_h \in X_h$ is now defined as the unique minimizer of the discrete energy functional $\mathcal{E}_h$ restricted to the subspace X_{hk} of X_h, i.e.we look for

$$u_{hk} = \operatorname*{argmin}_{x_{hk} \in X_{hk}} \mathcal{E}_h(x_{hk}) = \operatorname*{argmin}_{x_{hk} \in X_{hk}} \frac{1}{2} \left(\|\mathcal{A}_h x_{hk} - f_h\|_{Y_h'}^2 + \alpha \|x_{hk}(0) - u_{0h}\|_L^2 \right). \tag{27}$$

We emphasize the use of the space semi-discrete test space Y_h to measure the dual norm of the residual. We obtain the following quasi-optimal error estimate.

Lemma 3.3 (Error estimate)**.** *Let u_h be the unique solution to* (23)*, and let $u_{hk} \in X_{hk}$ be the unique minimizer of* (27)*. Then, we have*

$$\|u_h - u_{hk}\|_{X_h} \leq C \inf_{x_{hk} \in X_{hk}} \|u_h - x_{hk}\|_{X_h}, \tag{28}$$

where $C = \frac{\sqrt{3}M}{\beta}$ is independent of the time discretization.

Proof. Using (25), we have

$$\begin{aligned} \|u_{hk} - u_h\|_{X_h} &\leq \beta^{-1} \|\mathcal{B}_h(u_{hk} - u_h)\|_{Z_h'} = \beta^{-1} \min_{x_{hk} \in X_{hk}} \|\mathcal{B}_h(x_{hk} - u_h)\|_{Z_h'} \\ &\leq \sqrt{3} M \beta^{-1} \min_{x_{hk} \in X_{hk}} \|x_{hk} - u_h\|_{X_h}, \end{aligned}$$

which proves the assertion. □

The unique minimizer of the quadratic discrete minimization problem (27) can be characterized by a system of linear equations. To write this system, let us first introduce the Riesz isomorphism $R_{V_h'} : V_h \to V_h'$ such that $\langle R_{V_h'} v_h, w_h \rangle_{V_h', V_h} = \langle v_h, w_h \rangle_V$ for all $v_h, w_h \in V_h$ (note that $R_{V_h'}$ differs from the injection J_{V_h} introduced above). Let $R_{Y_h'} : Y_h \to Y_h'$ be the space-time Riesz isomorphism such that

$$R_{Y_h'} = R_{V_h'} \otimes I_{L^2}, \tag{29}$$

where I_{L^2} is the identity operator in $L^2(I)$ (it is actually the Riesz isomorphism from $L^2(I)$ onto $L^2(I)' \equiv L^2(I)$). The quadratic discrete minimization problem (27) is equivalent to the following linear problem: find $u_{hk} \in X_{hk}$ such that

$$B_{hk} u_{hk} = g_{hk}, \tag{30}$$

with $B_{hk} : X_{hk} \to X_{hk}'$ and $g_{hk} \in X_{hk}'$ such that, for all $x_{hk}, z_{hk} \in X_{hk}$,

$$\langle B_{hk} x_{hk}, z_{hk} \rangle_{X_{hk}', X_{hk}} = \langle \mathcal{A}_h x_{hk}, R_{Y_h'}^{-1} \mathcal{A}_h z_{hk} \rangle_{Y_h', Y_h} + \alpha \langle x_{hk}(0), z_{hk}(0) \rangle_L, \tag{31a}$$

$$\langle g_{hk}, z_{hk} \rangle_{X_{hk}', X_{hk}} = \langle f_h, R_{Y_h'}^{-1} \mathcal{A}_h z_{hk} \rangle_{Y_h', Y_h} + \alpha \langle u_{0h}, z_{hk}(0) \rangle_L. \tag{31b}$$

Let us briefly describe the algebraic realization of the discrete problem (30). Let $(\psi_i)_{1 \leq i \leq N_h}$ be a basis of V_h and let $(\phi_l)_{1 \leq l \leq N_k}$ be a basis of S_k. We can then seek for the components of the unique solution u_{hk} of (30) in the basis $(\psi_i \otimes \phi_l)_{1 \leq i \leq N_h, 1 \leq l \leq N_k}$ of X_{hk}, i.e., we seek $\mathbf{u} = (\mathbf{u}_{il})_{1 \leq i \leq N_h, 1 \leq l \leq N_k} \in \mathbb{R}^{N_h N_k}$ such that

$$u_{hk} = \sum_{i=1}^{N_h} \sum_{l=1}^{N_k} \mathbf{u}_{il} \psi_i \otimes \phi_l. \tag{32}$$

We define the following matrices of size $N_h \times N_h$ (related to the space discretization):

$$(\mathbf{D}_h)_{ij} = \langle \psi_j, \psi_i \rangle_V, \qquad (\mathbf{M}_h)_{ij} = \langle \psi_j, \psi_i \rangle_L, \tag{33}$$

and the following matrices of size $N_k \times N_k$ (related to the time discretization):

$$(\mathbf{D}_k)_{lm} = \int_I \phi'_m(t)\phi'_l(t)dt, \quad (\mathbf{M}_k)_{lm} = \int_I \phi_m(t)\phi_l(t)dt, \quad (\mathbf{O}_k)_{lm} = \phi_m(0)\phi_l(0). \tag{34}$$

Recalling the separated form (7), we introduce the following matrix of size $N_h \times N_h$:

$$(\mathbf{A}_h^{(p)})_{ij} = \langle A^{(p)}\psi_j, \psi_i \rangle_{V',V}, \tag{35}$$

and the following matrices of size $N_k \times N_k$:

$$(\mathbf{M}_k^{(p,p')})_{lm} = \int_I \mu^{(p)}(t)\mu^{(p')}(t)\phi_m(t)\phi_l(t)dt, \quad (\mathbf{E}_k^{(p)})_{lm} = \int_I \mu^{(p)}(t)\phi'_m(t)\phi_l(t)dt, \tag{36}$$

for all $1 \le p, p' \le P$. Then, we obtain the following symmetric positive-definite linear system in $\mathbb{R}^{N_h N_k}$:

$$\mathbf{B}\mathbf{u} = \mathbf{g}, \tag{37}$$

with the matrix

$$\begin{aligned}\mathbf{B} = \mathbf{M}_h\mathbf{D}_h^{-1}\mathbf{M}_h \otimes \mathbf{D}_k + \sum_{1\le p\le P} 2\mathrm{sym}\big\{(\mathbf{A}_h^{(p)})^{\mathrm{T}}\mathbf{D}_h^{-1}\mathbf{M}_h \otimes \mathbf{E}_k^{(p)}\big\} \\ + \sum_{1\le p,p'\le P} (\mathbf{A}_h^{(p)})^{\mathrm{T}}\mathbf{D}_h^{-1}\mathbf{A}_h^{(p')} \otimes \mathbf{M}_k^{(p,p')} + \alpha\mathbf{M}_h \otimes \mathbf{O}_k,\end{aligned} \tag{38}$$

where $\mathrm{sym}(\mathbf{Z}_h \otimes \mathbf{Z}_k) = \frac{1}{2}(\mathbf{Z}_h \otimes \mathbf{Z}_k + \mathbf{Z}_h^{\mathrm{T}} \otimes \mathbf{Z}_k^{\mathrm{T}})$ for any matrix $\mathbf{Z}_h$ of size $N_h \times N_h$ and any matrix $\mathbf{Z}_k$ of size $N_k \times N_k$, and the right-hand side

$$\mathbf{g} = \sum_{1\le q\le Q} \mathbf{M}_h\mathbf{D}_h^{-1}\mathbf{f}_h^{(q)} \otimes \mathbf{e}_k^{(q)} + \sum_{\substack{1\le p\le P\\ 1\le q\le Q}} (\mathbf{A}_h^{(p)})^{\mathrm{T}}\mathbf{D}_h^{-1}\mathbf{f}_h^{(q)} \otimes \mathbf{d}_k^{(p,q)} + \alpha\mathbf{u}_{\mathbf{0}h} \otimes \mathbf{i}_k, \tag{39}$$

with the vectors $(\mathbf{f}_h^{(q)})_i = \langle f^{(q)}|_{V_h}, \psi_i \rangle_{V',V} = \langle f^{(q)}, \psi_i \rangle_{V',V}$, $(\mathbf{u}_{\mathbf{0}h})_i = \langle u_{0h}, \psi_i \rangle_L = \langle u_0, \psi_i \rangle_L$, for all $1 \le i \le N_h$, and $(\mathbf{e}_k^{(q)})_l = \int_I \lambda^{(q)}(t)\phi'_l(t)dt$, $(\mathbf{d}_k^{(p,q)})_l = \int_I \mu^{(p)}(t)\lambda^{(q)}(t)\phi_l(t)dt$, $(\mathbf{i}_k)_l = \phi_l(0)$, for all $1 \le l \le N_k$.

Example 3.4 (Heat equation)**.** *Let us consider the heat equation where $P = 1$, $\mu^{(1)}(t) \equiv 1$ and $A^{(1)} = -\Delta$, and let us equip the space V with the H^1-seminorm so that $\langle v, w \rangle_V = \int_\Omega \nabla v(x) \cdot \nabla w(x) dx = \langle A^{(1)}v, w \rangle_{V',V}$. Then the expression of* $\mathbf{B}$ *simplifies as follows:*

$$\mathbf{B} = \mathbf{M}_h\mathbf{D}_h^{-1}\mathbf{M}_h \otimes \mathbf{D}_k + \mathbf{M}_h \otimes 2\mathrm{sym}(\mathbf{E}_k) + \mathbf{D}_h \otimes \mathbf{M}_k + \alpha\mathbf{M}_h \otimes \mathbf{O}_k, \tag{40}$$

with the following matrix of size $N_k \times N_k$:

$$(\mathbf{E}_k)_{lm} = \int_I \phi'_m(t)\phi_l(t)dt. \tag{41}$$

4 Low-rank approximation

In this section, we present the low-rank approximation method we use to approximate iteratively the unique minimizer of (27) (or, equivalently, the unique solution to the linear system (37)). We consider here a greedy algorithm [34, 5, 12, 7] which is an iterative procedure such that, at each iteration $m \in \mathbb{N}^*$, one computes an approximation $u_{hk}^m \in X_{hk}$ of the solution $u_{hk} \in X_{hk}$ of (27) in the form

$$u_{hk}^m(x,t) = \sum_{1 \leq n \leq m} v_h^n(x) \otimes s_k^n(t), \tag{42}$$

with $v_h^n \in V_h$ and $s_k^n \in S_k$ for all $1 \leq n \leq m$. The algorithm can be outlined as follows:

GREEDY ALGORITHM:

1. Set $u_{hk}^0 = 0$ and $m = 1$.
2. Solve for

$$(v_h^m, s_k^m) \in \underset{(v_h, s_k) \in V_h \times S_k}{\operatorname{argmin}} \mathcal{E}_h(u_{hk}^{m-1} + v_h \otimes s_k). \tag{43}$$

   Set $u_{hk}^m := u_{hk}^{m-1} + v_h^m \otimes s_k^m$.
3. Check convergence, and if not satisfied, set $m \leftarrow m+1$ and go to step (2).

The following relative stagnation-based stopping criterion is used with a tolerance $\epsilon_{\text{greedy}} > 0$:

$$\frac{\|v_h^m \otimes s_k^m\|_X}{\|u_{hk}^m\|_X} < \epsilon_{\text{greedy}}. \tag{44}$$

Using the general results from [5, 12], one can verify that the iterations of the above greedy algorithm are well-defined using the discrete minimal residual formulation presented in Section 3. Recall that the uniqueness of the solution to the minimization problem (27) follows from the strong convexity of the functional $\mathcal{E}_h$, and the sequence $(u_{hk}^m)_{m \in \mathbb{N}}$ converges to u_{hk} as n goes to infinity. Actually, it can be checked that this convergence result still holds true in the infinite-dimensional setting.

In the above greedy algorithm, the minimization problem (43) is nonlinear. Therefore, it is not straightforward to solve it and in practice, one often considers an alternating minimization algorithm (see [37]), based on the following fixed-point iterative scheme:

ALTERNATING MINIMIZATION ALGORITHM FOR (43):

1. Choose $s_k^{m,0} \in S_k$ randomly and set $p = 1$.

2. Let $v_h^{m,p} \in V_h$ be the unique solution to

$$v_h^{m,p} = \underset{v_h \in V_h}{\operatorname{argmin}}\ \mathcal{E}_h\left(u_{hk}^{m-1} + v_h \otimes s_k^{m,p-1}\right). \tag{45}$$

Compute $s_k^{m,p} \in S_k$ to be the unique solution to

$$s_k^{m,p} = \underset{s_k \in S_k}{\operatorname{argmin}}\ \mathcal{E}_h\left(u_{hk}^{m-1} + v_h^{m,p} \otimes s_k\right). \tag{46}$$

3. Check convergence, and if not satisfied, set $p \leftarrow p+1$ and go to step (2).

The following relative stagnation-based stopping criterion is used with a tolerance $\epsilon_{\text{alt}} > 0$:

$$\frac{\|v_h^{m,p} \otimes s_k^{m,p} - v_h^{m,p-1} \otimes s_k^{m,p-1}\|_X}{\|v_h^{m,p} \otimes s_k^{m,p}\|_X} < \epsilon_{\text{alt}}. \tag{47}$$

The cost of an iteration of the alternating minimization algorithm is of order $(N_h + N_k)$. Provided the number of fixed-point iterations remains reasonably small, the cost of each iteration of the greedy algorithm can be estimated to scale also as $(N_h + N_k)$. We will verify in our numerical experiments that this is indeed the case.

Remark 4.1 (Matrix form)**.** *The matrix form of problem (43) is as follows:*

$$(\mathbf{v}_h^m, \mathbf{s}_k^m) = \underset{(\mathbf{v}_h, \mathbf{s}_k) \in \mathbb{R}^{N_h} \times \mathbb{R}^{N_k}}{\operatorname{argmin}} \left\{ \frac{1}{2}\big(\mathbf{u}_{hk}^{m-1} + \mathbf{v}_h \otimes \mathbf{s}_k\big)^{\mathrm{T}} \mathbf{B} \big(\mathbf{u}_{hk}^{m-1} + \mathbf{v}_h \otimes \mathbf{s}_k\big) - \big(\mathbf{u}_{hk}^{m-1} + \mathbf{v}_h \otimes \mathbf{s}_k\big)^{\mathrm{T}} \mathbf{g} \right\},$$

where $\mathbf{u}_{hk}^{m-1}$ *denotes the vector in* $\mathbb{R}^{N_h N_k}$ *containing the coordinates of* u_{hk}^{m-1} *in the basis* $(\psi_i \otimes \phi_l)_{1 \le i \le N_h, 1 \le l \le N_k}$*. Similarly, the matrix form of problems* (45) *and* (46) *is as follows:*

$$\mathbf{v}_h^{m,p} = \underset{\mathbf{v}_h \in \mathbb{R}^{N_h}}{\operatorname{argmin}} \left\{ \frac{1}{2}\big(\mathbf{u}_{hk}^{m-1} + \mathbf{v}_h \otimes \mathbf{s}_k^{m,p-1}\big)^{\mathrm{T}} \mathbf{B} \big(\mathbf{u}_{hk}^{m-1} + \mathbf{v}_h \otimes \mathbf{s}_k^{m,p-1}\big) - \big(\mathbf{u}_{hk}^{m-1} + \mathbf{v}_h \otimes \mathbf{s}_k^{m,p-1}\big)^{\mathrm{T}} \mathbf{g} \right\},$$

$$\mathbf{s}_k^{m,p} = \underset{\mathbf{s}_k \in \mathbb{R}^{N_k}}{\operatorname{argmin}} \left\{ \frac{1}{2}\big(\mathbf{u}_{hk}^{m-1} + \mathbf{v}_h^{m,p} \otimes \mathbf{s}_k\big)^{\mathrm{T}} \mathbf{B} \big(\mathbf{u}_{hk}^{m-1} + \mathbf{v}_h^{m,p} \otimes \mathbf{s}_k\big) - \big(\mathbf{u}_{hk}^{m-1} + \mathbf{v}_h^{m,p} \otimes \mathbf{s}_k\big)^{\mathrm{T}} \mathbf{g} \right\}.$$

5 Other discrete minimal residual methods

In this section, we describe for the purpose of numerical comparisons in Section 6 two other discrete minimal residual approaches. The discrete method introduced

in Section 3 is henceforth referred to as Method 1. The first variant, called Method 2, hinges on a fully discrete Petrov–Galerkin setting as devised in [1, 2, 3]. The second variant, called Method 3, uses the same space semi-discrete setting as Method 1, but the test space is now equipped with a simple Euclidean norm of the components on a basis of V_h; Method 3 has been introduced in [30].

5.1 Method 2: fully discrete Petrov–Galerkin method

Let us set

$$Y_{hk} := V_h \otimes S_k^{\mathrm{P}}, \qquad Y'_{hk} = V'_h \otimes (S_k^{\mathrm{P}})', \tag{48}$$

where S_k^{P} is a finite-dimensional subspace of $L^2(I)$ so that $S_k^{\mathrm{P}} \subset L^2(I) \equiv L^2(I)' \subset (S_k^{\mathrm{P}})'$. The injection $J_{S_k^{\mathrm{P}}} : L^2(I) \to (S_k^{\mathrm{P}})'$ is such that $J_{S_k^{\mathrm{P}}} = R_{(S_k^{\mathrm{P}})'} \circ \Pi_{S_k^{\mathrm{P}}}$ where $\Pi_{S_k^{\mathrm{P}}}$ is the $L^2(I)$-orthogonal projection from $L^2(I)$ onto S_k^{P} and $R_{(S_k^{\mathrm{P}})'} : S_k^{\mathrm{P}} \to (S_k^{\mathrm{P}})'$ is the Riesz isomorphism so that $\langle R_{(S_k^{\mathrm{P}})'} q, r\rangle_{(S_k^{\mathrm{P}})', S_k^{\mathrm{P}}} = \langle q, r\rangle_{L^2(I)}$, for all $q, r \in S_k^{\mathrm{P}}$. Let us set $\dim(S_k^{\mathrm{P}}) = N_k^{\mathrm{P}}$. Recalling the separated form (7), let us define $f_{hk} \in Y'_{hk}$ s.t.

$$f_{hk}(t) = \sum_{1\le q\le Q} (J_{S_k^{\mathrm{P}}} \lambda^{(q)})(t) f^{(q)}|_{V_h} \qquad \text{a.e. } t \in I. \tag{49}$$

Let $\mathcal{A}_{hk} := (I_{V'_h} \otimes J_{S_k^{\mathrm{P}}})\mathcal{A}_h : X_{hk} \to Y'_{hk}$ where $I_{V'_h}$ is the identity operator in V'_h and $\mathcal{A}_h$ is defined by (20). We consider the discrete energy functional $\mathcal{E}_{hk}^{\mathrm{fdPG}} : X_{hk} \to \mathbb{R}$ defined as

$$\mathcal{E}_{hk}^{\mathrm{fdPG}}(x_{hk}) := \frac{1}{2}\left(\|\mathcal{A}_{hk}x_{hk} - f_{hk}\|_{Y'_{hk}}^2 + \alpha\|x_{hk}(0) - u_{0h}\|_L^2\right), \qquad \forall x_{hk} \in X_{hk}. \tag{50}$$

The discrete minimization problem is as follows: find $u_{hk}^{\mathrm{fdPG}} \in X_{hk}$ such that

$$u_{hk}^{\mathrm{fdPG}} = \operatorname*{argmin}_{x_{hk}\in X_{hk}} \mathcal{E}_{hk}^{\mathrm{fdPG}}(x_{hk}). \tag{51}$$

Remark 5.1 (Comparison of energies)**.** *Since $f_{hk} = (I_{V'_h} \otimes J_{S_k^{\mathrm{P}}})f_h$ with f_h defined by* (19)*, we have*

$$\|\mathcal{A}_{hk}x_{hk} - f_{hk}\|_{Y'_{hk}} = \|(I_{V'_h} \otimes J_{S_k^{\mathrm{P}}})(\mathcal{A}_h x_{hk} - f_h)\|_{Y'_{hk}} \le \|\mathcal{A}_h x_{hk} - f_h\|_{Y'_h},$$

which implies that $\mathcal{E}_{hk}^{\mathrm{fdPG}}(x_{hk}) \le \mathcal{E}_h(x_{hk})$ for all $x_{hk} \in X_{hk}$.

As shown in [2, 3] in the case of time-independent and selfadjoint operators $A \in \mathcal{L}(V; V')$, the Hessian of the discrete energy $\mathcal{E}_{hk}^{\mathrm{fdPG}}$ induces a bilinear form that satisfies an inf-sup condition that degenerates with the parabolic CFL. In the general case with a time-dependent differential operator, the positivity of the inf-sup constant is not guaranteed a priori, which means that the discrete energy functional $\mathcal{E}_{hk}^{\mathrm{fdPG}}$ may be only convex in some unfavorable situations. This means that in such cases, global minimizers of (51) exist but may not be unique. Any minimizer satisfies the following system of linear equations: find $u_{hk}^{\mathrm{fdPG}} \in X_{hk}$ such that

$$B_{hk}^{\mathrm{fdPG}} u_{hk}^{\mathrm{fdPG}} = g_{hk}^{\mathrm{fdPG}}, \tag{52}$$

with $B_{hk}^{\mathrm{fdPG}} : X_{hk} \to X'_{hk}$ and $g_{hk}^{\mathrm{fdPG}} \in X'_{hk}$ such that, for all $x_{hk}, z_{hk} \in X_{hk}$,

$$\langle B_{hk}^{\mathrm{fdPG}} x_{hk}, z_{hk}\rangle_{X'_{hk},X_{hk}} = \langle \mathcal{A}_{hk} x_{hk}, R_{Y'_{hk}}^{-1} \mathcal{A}_{hk} z_{hk}\rangle_{Y'_{hk},Y_{hk}} + \alpha \langle x_{hk}(0), z_{hk}(0)\rangle_L, \tag{53a}$$

$$\langle g_{hk}^{\mathrm{fdPG}}, z_{hk}\rangle_{X'_{hk},X_{hk}} = \langle f_{hk}, R_{Y'_{hk}}^{-1} \mathcal{A}_{hk} z_{hk}\rangle_{Y'_{hk},Y_{hk}} + \alpha \langle u_{0h}, z_{hk}(0)\rangle_L, \tag{53b}$$

with the space-time Riesz isomorphism $R_{Y'_{hk}} = R_{V'_h} \otimes J_{S_k^{\mathrm{P}}} : Y_{hk} \to Y'_{hk}$. Let us point out that, since $J_{S_k^{\mathrm{P}}} = R_{(S_k^{\mathrm{P}})'}$ on S_k^{P}, we have $R_{Y'_{hk}} = R_{V'_h} \otimes R_{(S_k^{\mathrm{P}})'}$. Let us briefly describe the algebraic realization of the discrete problem (52). Recall that $(\psi_i)_{1\le i\le N_h}$ is a basis of V_h and $(\phi_l)_{1\le l\le N_k}$ is a basis of S_k. Let $(\phi_l^{\mathrm{P}})_{1\le l\le N_k^{\mathrm{P}}}$ be a basis of S_k^{P}. In addition to the square matrices $\mathbf{M}_h$, $\mathbf{D}_h$, $\mathbf{O}_k$ defined by (33), (34), (35), we consider the square matrix $\mathbf{M}_k^{\mathrm{P}}$ of size $N_k^{\mathrm{P}} \times N_k^{\mathrm{P}}$ and the rectangular matrix $\mathbf{E}_k^{\mathrm{PG}}$ of size $N_k^{\mathrm{P}} \times N_k$ such that

$$(\mathbf{M}_k^{\mathrm{P}})_{lm} = \int_I \phi_m^{\mathrm{P}}(t)\phi_l^{\mathrm{P}}(t)dt, \qquad (\mathbf{E}_k^{\mathrm{PG}})_{lm} = \int_I \phi'_m(t)\phi_l^{\mathrm{P}}(t)dt, \tag{54}$$

and the rectangular matrices $\mathbf{M}_k^{\mathrm{PG,(p)}}$, for all $1 \le p \le P$, of size $N_k^{\mathrm{P}} \times N_k$ such that

$$(\mathbf{M}_k^{\mathrm{PG,(p)}})_{lm} = \int_I \mu^{(p)}(t)\phi_m(t)\phi_l^{\mathrm{P}}(t)dt. \tag{55}$$

Then, we obtain the following symmetric positive-definite linear system in $\mathbb{R}^{N_h N_k}$:

$$\mathbf{B}^{\mathrm{fdPG}}\mathbf{u}^{\mathrm{fdPG}} = \mathbf{g}^{\mathrm{fdPG}}, \tag{56}$$

with the matrix

$$\begin{aligned}\mathbf{B}^{\mathrm{fdPG}} = {}& \mathbf{M}_h\mathbf{D}_h^{-1}\mathbf{M}_h \otimes (\mathbf{E}_k^{\mathrm{PG}})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{E}_k^{\mathrm{PG}} \\ & + \sum_{1\le p\le P} 2\mathrm{sym}\big\{(\mathbf{A}_h^{(p)})^{\mathrm{T}}\mathbf{D}_h^{-1}\mathbf{M}_h \otimes (\mathbf{M}_k^{\mathrm{PG,(p)}})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{E}_k^{\mathrm{PG}}\big\} \\ & + \sum_{1\le p,p'\le P} (\mathbf{A}_h^{(p)})^{\mathrm{T}}\mathbf{D}_h^{-1}\mathbf{A}_h^{(p')} \otimes (\mathbf{M}_k^{\mathrm{PG,(p)}})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{M}_k^{\mathrm{PG,(p')}} + \alpha\mathbf{M}_h \otimes \mathbf{O}_k,\end{aligned} \tag{57}$$

and the right-hand side

$$\begin{aligned}\mathbf{g}^{\mathrm{fdPG}} = {}& \sum_{1\le q\le Q} \mathbf{M}_h\mathbf{D}_h^{-1}\mathbf{f}_h^{(q)} \otimes (\mathbf{E}_k^{\mathrm{PG}})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{e}_k^{\mathrm{P,(q)}} \\ & + \sum_{\substack{1\le p\le P\\ 1\le q\le Q}} (\mathbf{A}_h^{(p)})^{\mathrm{T}}\mathbf{D}_h^{-1}\mathbf{f}_h^{(q)} \otimes (\mathbf{M}_k^{\mathrm{PG,(p)}})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{d}_k^{\mathrm{P,(q)}} \\ & + \alpha\mathbf{u}_{\mathbf{0}h} \otimes \mathbf{i}_k,\end{aligned} \tag{58}$$

with $(\mathbf{e}_k^{\mathrm{P,(q)}})_l = \int_I \lambda^{(q)}(t)(\phi_l^{\mathrm{P}})'(t)dt$ and $(\mathbf{d}_k^{\mathrm{P,(q)}})_l = \int_I \lambda^{(q)}(t)\phi_l^{\mathrm{P}}(t)dt$, for all $1 \le l \le N_k^{\mathrm{P}}$.

Remark 5.2 (Lowest-order Petrov–Galerkin discretization)**.** *Assume that S_k is composed of continuous, piecewise affine functions and that S_k^{P} is composed of piecewise constant functions on the same time mesh so that* $\dim(S_k^{\mathrm{P}}) = N_k - 1$.

This corresponds to the well-known Crank–Nicolson time scheme. Then, one can readily verify that

$$(\mathbf{E}_k^{\mathrm{PG}})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{E}_k^{\mathrm{PG}} = \mathbf{D}_k, \qquad (\mathbf{M}_k^{\mathrm{PG},(\mathrm{p})})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{E}_k^{\mathrm{PG}} = \mathbf{E}_k^{(p)}, \tag{59}$$

for all $1 \le p \le P$*, where* $\mathbf{D}_k$ *is defined in* (34) *and* $\mathbf{E}_k^{(p)}$ *in* (36)*. As a consequence, there is only one term composing the matrices* $\mathbf{B}$ *and* $\mathbf{B}^{\mathrm{fdPG}}$ *that differs, namely the time matrix in the double summation over* p, p' *where this matrix is* $\mathbf{M}_k^{(p,p')}$ *for* $\mathbf{B}$ *and is* $(\mathbf{M}_k^{\mathrm{PG},(\mathrm{p})})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{M}_k^{\mathrm{PG},(\mathrm{p}')}$ *for* $\mathbf{B}^{\mathrm{fdPG}}$*. Even for the heat equation with* $P = 1$ *and* $\mu^{(1)}(t) \equiv 1$*, these matrices, which become, respectively,* $\mathbf{M}_k$ *and* $(\mathbf{M}_k^{\mathrm{PG}})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{M}_k^{\mathrm{PG}}$ *with* $(\mathbf{M}_k^{\mathrm{PG}})_{lm} = \int_I \phi_m(t)\phi_l^{\mathrm{P}}(t)dt$*, are still different. Note that* $\mathbf{M}_k \ge (\mathbf{M}_k^{\mathrm{PG}})^{\mathrm{T}}(\mathbf{M}_k^{\mathrm{P}})^{-1}\mathbf{M}_k^{\mathrm{PG}}$ *in the sense of quadratic forms, which is compatible with our above observation on the discrete energies that* $\mathcal{E}_{hk}^{\mathrm{fdPG}}(x_{hk}) \le \mathcal{E}_h(x_{hk})$ *for all* $x_{hk} \in X_{hk}$*. As observed in [1, 2, 3], uniform stability with respect to the time discretization is not guaranteed, but this can be fixed, e.g., by constructing the discrete test space* S_k^{P} *using a time mesh that is twice as fine as that used for the discrete trial space.*

5.2 Method 3: an unpreconditioned space semi-discrete Galerkin method

We consider the same space semi-discrete setting as in Section 3 but we now equip the space V_h with the Euclidean norm of the components on the basis $(\psi_i)_{1\le i\le N_h}$ instead of considering as before the norm induced by V. The main motivation for this change is that it avoids the appearance of the inverse stiffness matrix $\mathbf{D}_h^{-1}$ in the linear system. We obtain the following symmetric positive-definite linear system in $\mathbb{R}^{N_h N_k}$:

$$\mathbf{B}^{\mathrm{unprec}}\mathbf{u}^{\mathrm{unprec}} = \mathbf{g}^{\mathrm{unprec}}, \tag{60}$$

with the matrix

$$\begin{aligned}\mathbf{B}^{\mathrm{unprec}} = (\mathbf{M}_h\mathbf{I}_h\mathbf{M}_h) \otimes \mathbf{D}_k + \sum_{1\le p\le P} 2\mathrm{sym}\big\{((\mathbf{A}_h^{(p)})^{\mathrm{T}}\mathbf{I}_h\mathbf{M}_h) \otimes \mathbf{E}_k^{(p)}\big\} \\ + \sum_{1\le p,p'\le P} ((\mathbf{A}_h^{(p)})^{\mathrm{T}}\mathbf{I}_h\mathbf{A}_h^{(p')}) \otimes \mathbf{M}_k^{(p,p')} + \alpha\mathbf{M}_h \otimes \mathbf{O}_k,\end{aligned} \tag{61}$$

and the right-hand side

$$\mathbf{g}^{\mathrm{unprec}} = \sum_{1\le q\le Q} \mathbf{M}_h\mathbf{I}_h\mathbf{f}_h^{(q)} \otimes \mathbf{e}_k^{(q)} + \sum_{\substack{1\le p\le P\\ 1\le q\le Q}} (\mathbf{A}_h^{(p)})^{\mathrm{T}}\mathbf{I}_h\mathbf{f}_h^{(q)} \otimes \mathbf{d}_k^{(p,q)} + \alpha\mathbf{u}_{\mathbf{0}h} \otimes \mathbf{i}_k, \tag{62}$$

where $\mathbf{I}_h$ is the identity matrix of size $N_h \times N_h$. The present formulation is chosen for illustrative purposes; in practice, one can also replace $\mathbf{D}_h^{-1}$ by another matrix.

6 Numerical results

For all the test cases, we consider the space domain $\Omega = (0,1) \times (0,1)$, the time interval $I = [0,1]$, and the functional spaces $V = H_0^1(\Omega)$ and $L = L^2(\Omega)$.

We consider first the heat equation, where the differential operator A is time-independent and selfadjoint, then we consider a time-oscillatory diffusion problem, where the operator is time-dependent and selfadjoint, and finally a convection-diffusion equation, where the operator is time-independent and non-selfadjoint. The scaling factor for the contribution of the initial condition to the residual functional is always taken to be $\alpha := 1$. Let $\mathcal{T}_h$ be a shape-regular mesh of the domain Ω; in what follows, we consider uniform meshes composed of square cells. The finite-dimensional finite element subspace $V_h \subset V$ of dimension N_h is composed of continuous, piecewise bilinear functions on $\mathcal{T}_h$ vanishing at the boundary. Let $\mathcal{T}_k$ be a mesh of the interval I; for simplicity, we consider uniform meshes in time. The finite-dimensional subspace $S_k \subset H^1(I)$ of dimension N_k is composed of continuous, piecewise affine functions on $\mathcal{T}_k$. In what follows, all the norms of residuals of algebraic quantities are evaluated using the Euclidean norm in $\mathbb{R}^{N_h N_k}$ which is denoted $\|\cdot\|_{\ell^2}$. When comparing to Method 2 (see Section 5.1), we considered the Crank–Nicolson time scheme discussed in Remark 5.2. We also performed systematic comparisons with the uniformly stable variant using a finer time mesh for the test space, but we did not observe any significant difference in the results obtained for all the test cases considered herein.

6.1 Test case 1: heat equation with manufactured solution

We consider the heat equation with the time-independent, selfadjoint operator $A = -\Delta$. The initial condition is zero and the source term is evaluated from the following manufactured solution:

$$u(x, y, t) = \sum_{1 \le n \le 10} n^{-4} \sin(\pi n^3 t) \sin(\pi n x) \sin(\pi n y). \tag{63}$$

The discretization parameters are $N_h = (2^6)^2$ and $N_k = 2^{13}$, and the stopping tolerances are $\epsilon_{\text{greedy}} = 10^{-5}$ and $\epsilon_{\text{alt}} = 5 \times 10^{-2}$.

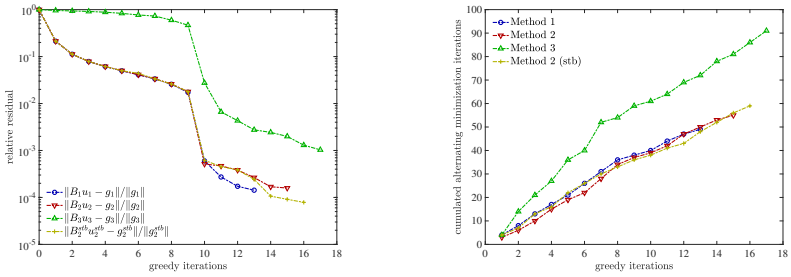

Figure 1: Test case 1. Left: relative residual at each iteration of the greedy algorithm. Right: cumulated number of alternating minimization iterations in the greedy algorithm.

The left panel of Figure 1 presents the decrease of the relative residual as a function of the number of greedy iterations for Methods 1, 2 and 3. More precisely, we plot $\|\mathbf{B}_i \mathbf{u}_i^m - \mathbf{g}_i\|_{\ell^2} / \|\mathbf{g}_i\|_{\ell^2}$ where $i \in \{1, 2, 3\}$ is the method index and m is the greedy iteration counter. We notice that the three methods

take about the same number of iterations (13, 15, and 17, respectively). This number is slightly larger than the space-time rank of the manufactured exact solution which is equal to 10. However, the relative residual takes larger values for Method 3 than for Methods 1 and 2. The right panel of Figure 1 presents the cumulated number of alternating minimization iterations in the greedy algorithm for Methods 1, 2 and 3. We observe that this number is about the same for Methods 1 and 2, whereas it is about 1.8 times larger for Method 3. Therefore, the use of a preconditioner, although it requires some additional computational effort, is beneficial to the efficiency of the overall behavior of the greedy algorithm. It is interesting to notice that with Methods 1 and 2, we have solved at convergence of the greedy algorithm about 50 linear systems in space, which is about 0.25% of the amount that would have been solved by using an implicit time-stepping method (recall that $N_k = 2^{13}$). We can make two further remarks concerning the decrease of the (relative) residual. First, we can decompose the residual of the space-time linear system as follows:

$$\mathbf{B}_i \mathbf{u}_i^m - \mathbf{g}_i = (\mathbf{B}_i^{\text{pde}} \mathbf{u}_i^m - \mathbf{g}_i^{\text{pde}}) + (\mathbf{B}_i^{\text{ic}} \mathbf{u}_i^m - \mathbf{g}_i^{\text{ic}}), \tag{64}$$

where we have written $\mathbf{B}_i = \mathbf{B}_i^{\text{pde}} + \mathbf{B}_i^{\text{ic}}$ and $\mathbf{g}_i = \mathbf{g}_i^{\text{pde}} + \mathbf{g}_i^{\text{ic}}$ to distinguish the contribution of the differential operator from that of the initial condition. Our results (not displayed for brevity) show that after a few greedy iterations, the two contributions have about the same size. Moreover, as the greedy iteration approaches convergence, there is some compensation between the two contributions to the relative residual in Method 1 (but not for Method 2) since they have a size which is about one order of magnitude larger than the relative residual itself. As a further comparison, we considered the quantities

$$r_i^m = \|\mathbf{B}_1 \mathbf{u}_i^m - \mathbf{g}_1\|_{\ell^2} / \|\mathbf{g}_1\|_{\ell^2}, \tag{65}$$

which represent the relative residual for the linear system originating from Method 1 when the iterates produced by Method $i \in \{1, 2, 3\}$ are inserted into the residual. As expected from the MinRes formulation, $r_1^m \leq \min(r_2^m, r_3^m)$ for all $m \geq 0$, and as the greedy iterations approach convergence, r_1^m reaches the value 4×10^{-5}, whereas r_2^m and r_3^m reach a value of 4×10^{-4} and 10^{-4}, respectively.

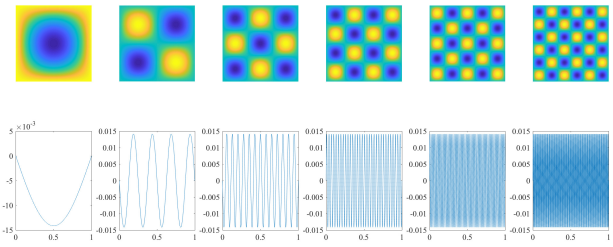

Figure 2: Test case 1: first six modes in space (top row) and in time (bottom row) for Method 1.

Figure 2 presents the first six space and time modes for Method 1. The first six modes obtained with Method 2 are essentially the same, whereas some

differences, especially in the space modes, can be observed with Method 3. This indicates that the preconditioner plays a relevant role in the exploration of the discrete trial space.

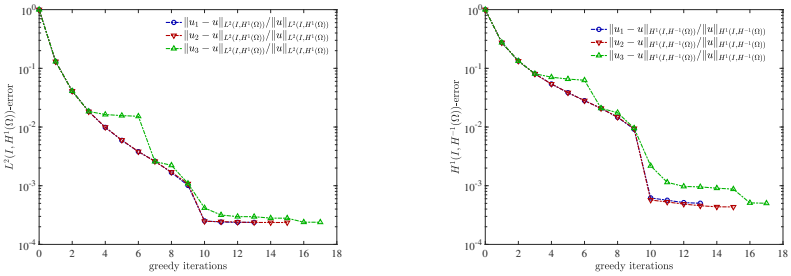

Figure 3: Test case 1: comparison of the errors produced by Methods 1, 2, 3 in two norms: $L^2(I;H^1(\Omega))$ (left) and $H^1(I;H^{-1}(\Omega))$ (right); in both cases, the curves for Methods 1 and 2 almost overlap.

Figure 3 reports the normalized errors $(u^m_{hk,i}-u)$ as a function of the iteration counter m of the greedy algorithm where, as above, the additional subscript $i \in \{1,2,3\}$ indicates which method has been used. The errors are measured in the $L^2(I;H^1(\Omega))$- and $H^1(I;H^{-1}(\Omega))$-norms. The three methods produce in both norms relatively close errors, and the error for Method 3 is always a bit larger. At convergence of the greedy algorithm, both errors are very small, namely 2×10^{-4} in the $L^2(I;H^1(\Omega))$-norm and 5×10^{-4} in the $H^1(I;H^{-1}(\Omega))$-norm.

Figure 4 presents a convergence study for Methods 1, 2 and 3 as a function of the discretization parameters N_h (in space) and N_k (in time). In both cases, we report the $L^2(I;H^1(\Omega))$- and $H^1(I;H^{-1}(\Omega))$-errors. The left panel considers $N_h = (2^l)^2$, $l \in \{2,3,4\}$ with $N_k = 2^{13}$, whereas the right panel considers $N_k = 2^l$, $l \in \{4,\dots,11\}$ with $N_h = (2^6)^2$. We observe that the convergence is of second order in time (if the time-step is small enough) and first order in space in the $L^2(I;H^1(\Omega))$-norm, whereas it is of first order in time (if the time-step is small enough) and second order in space in the $H^1(I;H^{-1}(\Omega))$-norm. These convergence orders are consistent with the expected decay rates of the best-approximation errors in both norms when approximating smooth functions by elements of the discrete trial space X_{hk}. Incidentally, we observe that the errors produced by Method 2 in both norms are slightly worse for the coarser time discretizations; this observation is consistent with the CFL-dependent inf-sup stability estimate for Method 2.

6.2 Test case 2: time-dependent diffusion

We consider a time-dependent, selfadjoint differential operator $A(t) = -\mu(t)\Delta$ with diffusion coefficient $\mu(t) = \sin(100\pi t) + 2$. The initial condition is $u_0 = 0$ and the source term is $f = 1$. The explicit expression of the exact solution is not available. The discretization parameters are $N_h = (2^6)^2$ and $N_k = 2^{13}$ (as in the previous test case), and the stopping tolerances are $\epsilon_{\text{greedy}} = 10^{-5}$ and $\epsilon_{\text{alt}} = 5\times 10^{-2}$ (as in the previous test case).

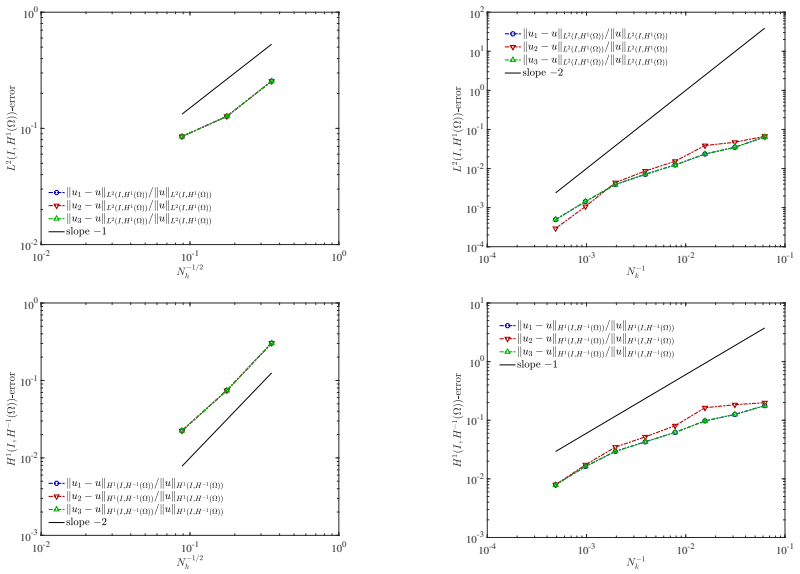

Figure 4: Test case 1: convergence study for Methods 1, 2 and 3 for errors measured in the $L^2(I;H^1(\Omega))$-norm (top row) and in the $H^1(I;H^{-1}(\Omega))$-norm (bottom row) for various mesh-sizes $N_h^{-1/2}$ (left column) and various time-steps N_k^{-1} (right column); in all cases, the curves for Methods 1 and 3 overlap.

The left panel of Figure 5 presents the decrease of the relative residual as a function of the number of greedy iterations for Methods 1, 2 and 3. We notice that the greedy algorithm takes between 16 and 21 iterations for the three methods to converge. The right panel of Figure 5 presents the cumulated number of alternating minimization iterations in the greedy algorithm for Methods 1, 2 and 3. We observe that this number is about the same for Methods 1 and 2 (as for test case 1), whereas it is about 1.5 times larger for Method 3, confirming once again the benefit of using a preconditioner. It is interesting to notice that with Methods 1 and 2, we have solved at convergence of the greedy algorithm about 80 linear systems in space, which is about 0.5% of the amount that would have been solved by using an implicit time-stepping method (recall that $N_k = 2^{13}$). Furthermore, similar observations as for test cases 1 and 2 can be made concerning the two contributions to the relative residual and the behavior of the residuals r_i^m defined by (65). In particular, we have again $r_1^m \leq \min(r_2^m, r_3^m)$ for all $m \geq 0$ (as expected from the MinRes formulation); as the greedy algorithm approaches convergence, r_1^m reaches a value of 9×10^{-5}, whereas r_2^m and r_3^m reach a value of 10^{-4} and 2×10^{-4}, respectively.

Figure 6 presents the first six space and time modes for Method 1. The first six modes obtained with Method 2 are essentially the same, whereas some differences, especially in the space modes, can be observed with Method 3. This observation again confirms that the preconditioner plays a relevant role in the exploration of the discrete trial space.

Figure 7 reports the normalized differences $(u_{hk,1}^m - u_{hk,2}^m)$ and $(u_{hk,1}^m - u_{hk,3}^m)$ as a function of the iteration counter m of the greedy algorithm, where, as

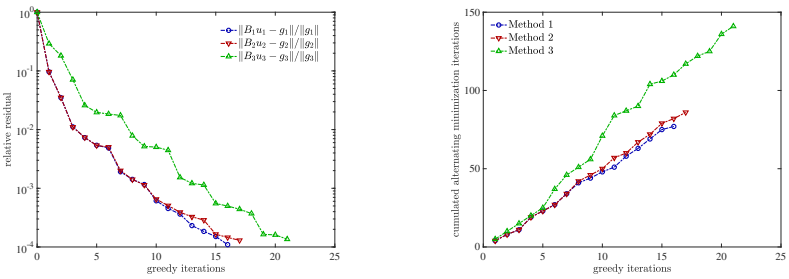

Figure 5: Test case 2. Left: relative residual at each iteration of the greedy algorithm. Right: cumulated number of alternating minimization iterations in the greedy algorithm.

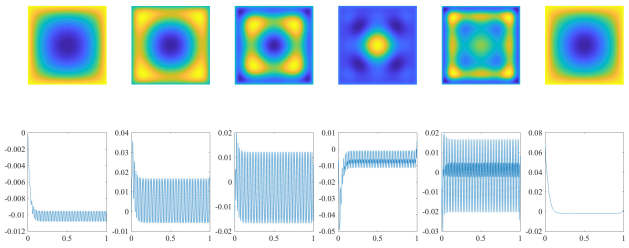

Figure 6: Test case 2: first six modes in space (top row) and in time (bottom row) for Method 1.

above, the additional subscript $i \in \{1, 2, 3\}$ indicates which method has been used. These differences are measured in the $L^2(I; H^1(\Omega))$- and $H^1(I; H^{-1}(\Omega))$-norms. We observe that the three methods produce approximate solutions that are relatively close in both norms. At the convergence of the greedy algorithm, the difference in the $L^2(I; H^1(\Omega))$-norm is of the order of 5×10^{-5}, and it is about one order of magnitude higher in the $H^1(I; H^{-1}(\Omega))$-norm.

Figure 8 presents a convergence study for Methods 1, 2 and 3 as a function of the discretization parameters N_h (in space) and N_k (in time). In both cases, we report the $L^2(I; H^1(\Omega))$- and $H^1(I; H^{-1}(\Omega))$-errors. The left panel considers $N_h = (2^l)^2$, $l \in \{2, 3, 4\}$ with $N_k = 2^{13}$, whereas the right panel considers $N_k = 2^l$, $l \in \{4, \ldots, 10\}$ with $N_h = (2^6)^2$. Since the exact solution is not available, we consider for each method the approximate solution produced on the finest space-time discretization available. These method-dependent reference solutions are very close according to Figure 7, and their differences are, in both norms, two orders of magnitude lower than the convergence errors reported in Figure 8. In this figure, we observe that for the three methods, the convergence rates are consistent with the best-approximation properties of the discrete trial space X_{hk} in both norms: the convergence is of second order in time and first order in space in the $L^2(I; H^1(\Omega))$-norm, whereas it is of first order in time and second order in space in the $H^1(I; H^{-1}(\Omega))$-norm.

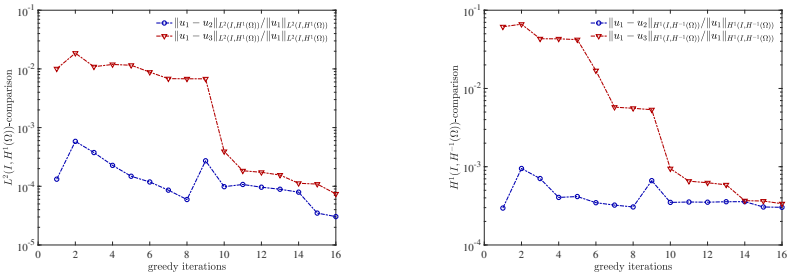

Figure 7: Test case 2: comparison of the solutions produced by Methods 1, 2, 3 in two norms: $L^2(I;H^1(\Omega))$ (left) and $H^1(I;H^{-1}(\Omega))$ (right).

6.3 Test case 3: advection-diffusion

We consider in this section a time-independent, but non-selfadjoint, differential operator $A = -\nabla\cdot\mu\nabla + c(x,y)\cdot\nabla$ with diffusion coefficient $\mu = 0.1$ and advection velocity field $c(x,y) = 2\pi(\frac{1}{2}-y, x-\frac{1}{2})^{\mathrm{T}}$. The source term is $f = 0$ and the initial condition is $u_0(x,y) = \exp\Big(-\frac{(x-\frac{2}{3})^2+(y-\frac{1}{2})^2}{0.07^2}\Big)$. The explicit expression of the exact solution is not available. The discretization parameters are $N_h = (2^5)^2$ and $N_k = 2^{10}$, and the stopping tolerances are $\epsilon_{\text{greedy}} = 10^{-5}$ and $\epsilon_{\text{alt}} = 5\times 10^{-2}$ (as in the previous test cases). The discretization parameters for this test case are a bit coarser than for the two other test cases because this test case turns out to be more computationally intensive. This is due to the fact that the differential operator in space is non-selfadjoint so that it is necessary to assemble the matrix $\mathbf{A}_h^{\mathrm{T}}\mathbf{D}_h^{-1}$ appearing in the definition (38) of the global system matrix $\mathbf{B}$ (recall that $P = 1$ here and that $\mathbf{A}_h^{\mathrm{T}} = \mathbf{D}_h$ when the differential operator corresponds to a pure diffusion operator).

The left panel of Figure 9 presents the decrease of the relative residual as a function of the number of greedy iterations for Methods 1, 2 and 3. We notice that for Methods 1 and 2, the greedy algorithm takes around 90 iterations to converge (94 and 97, respectively), whereas it takes only 49 iterations for Method 3. Thus, for this test case, Method 3 takes less iterations. The right panel of Figure 9 presents the cumulated number of alternating minimization iterations in the greedy algorithm for Methods 1, 2 and 3. We observe that this number is about the same for the three methods. When reaching convergence for the greedy algorithm, we have solved about 750 linear systems in space, which is 73% of the amount that would have been solved by using an implicit time-stepping method (recall that $N_k = 2^{10}$). This percentage is larger than the ones reported for the previous two test cases, but is still competitive. Furthermore, similar observations as for test cases 1 and 2 can be made concerning the two contributions to the relative residual and the behavior of the residuals r_i^m defined by (65). In particular, we have again $r_1^m \le \min(r_2^m, r_3^m)$ for all $m \ge 0$ (as expected from the MinRes formulation); as the greedy algorithm approaches convergence, r_1^m reaches a value of 2×10^{-5}, whereas r_2^m and r_3^m reach a value of 7×10^{-5} and 3×10^{-3}, respectively.

Figure 10 presents the first six space and time modes for Method 1. The first six modes obtained with Method 2 are essentially the same, whereas some

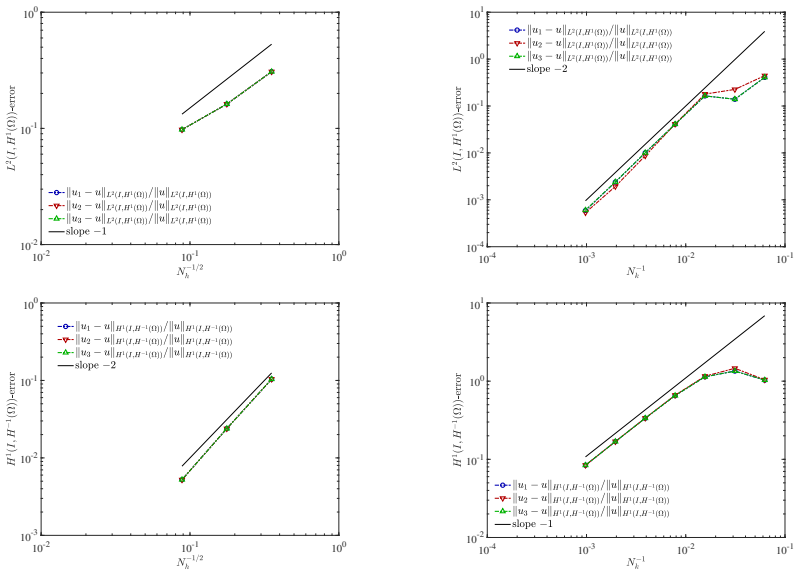

Figure 8: Test case 2: convergence study for Methods 1, 2 and 3 for errors measured in the $L^2(I;H^1(\Omega))$-norm (top row) and in the $H^1(I;H^{-1}(\Omega))$-norm (bottom row) for various mesh-sizes $N_h^{-1/2}$ (left column) and various time-steps N_k^{-1} (right column); in all cases, the curves for Methods 1 and 3 overlap.

differences, especially in the space modes, can be observed with Method 3. This observation again confirms that the preconditioner plays a relevant role in the exploration of the discrete trial space.

Figure 11 reports the normalized differences $(u^m_{hk,1}-u^m_{hk,2})$ and $(u^m_{hk,1}-u^m_{hk,3})$ as a function of the iteration counter m of the greedy algorithm where, as above, the additional subscript $i \in \{1,2,3\}$ indicates which method has been used. These differences are measured in the $L^2(I;H^1(\Omega))$- and $H^1(I;H^{-1}(\Omega))$-norms. We observe that the difference between the solutions produced by Methods 1 and 3 is significant in both norms (two orders of magnitude larger than the difference between Methods 1 and 2); therefore, we can conclude that Method 3 converges more rapidly than Methods 1 and 2, but with a poorer accuracy.

Figure 12 presents a convergence study for Methods 1, 2 and 3 as a function of the discretization parameters N_h (in space) and N_k (in time). In both cases, we report the $L^2(I;H^1(\Omega))$- and $H^1(I;H^{-1}(\Omega))$-errors. The left panel considers $N_h = (2^l)^2$, $l \in \{2,3,4\}$ with $N_k = 2^{10}$, whereas the right panel considers $N_k = 2^l$, $l \in \{4,\dots,8\}$ with $N_h = (2^5)^2$. Since the exact solution is not available, we consider for each method the approximate solution produced on the finest space-time discretization available. For Methods 1 and 2, these two reference solutions are very close according to Figure 11, and their difference is, in both norms, two orders of magnitude lower than the convergence errors reported in Figure 12. Moreover, for both methods, the reported convergence rates are, as above, consistent with the best-approximation properties of the discrete trial space X_{hk} in both norms. Finally, for Method 3, the convergence rates are similar except for the behavior with respect to time refinement in the

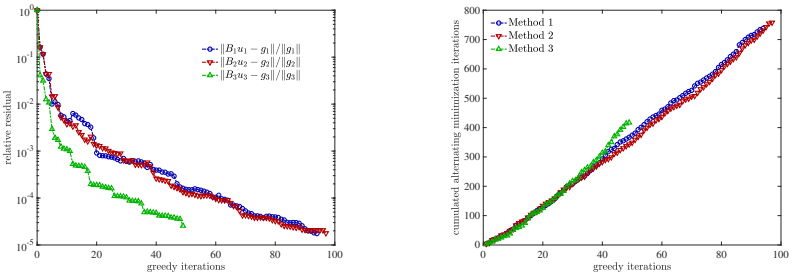

Figure 9: Test case 3. Left: relative residual at each iteration of the greedy algorithm. Right: cumulated number of alternating minimization iterations in the greedy algorithm.

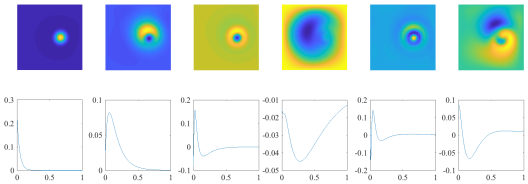

Figure 10: Test case 3: first six modes in space (top row) and in time (bottom row) for Method 1.

$L^2(I; H^1(\Omega))$-norm which is somewhat sub-optimal.

7 Conclusion and outlook

In this work, we have devised a space-time tensor for the low-rank approximation of linear parabolic evolution equations. The proposed method is uniformly stable with respect to the space-time discretization parameters and leads to solving sequentially global problems in space and in time. Our numerical results on various test cases indicate the importance of the preconditioner (which enters the method by means of the norm equipping the discrete test space), since the preconditioner has an influence on the convergence of the greedy algorithm. However, although the preconditioner improves the convergence of the greedy algorithm, it increases the cost of each iteration. Table 1 summarizes the relative CPU times for Methods 2 and 3 normalized with respect to Method 1 (we have considered 21 random initializations in each case and used for each method the median CPU time). We can see from this table that Methods 1 and 2 essentially deliver the same CPU times, whereas Method 3 turns out to be more effective especially for test case 3 despite the increased number of greedy iterations. Finally, various perspectives of this work can be envisaged. We mention in particular the question of adapting the discretization spaces and that of devising different approaches to obtain a separated representation of the exact solution with sufficient accuracy at low-rank when the differential operator has

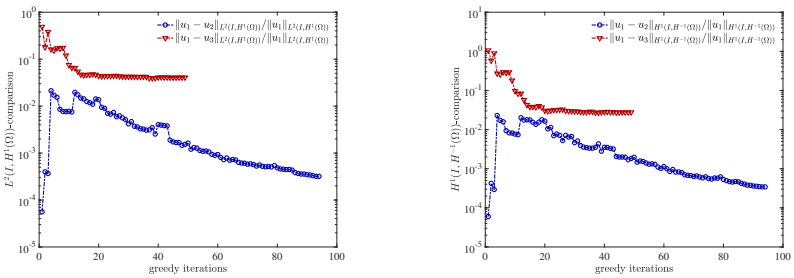

Figure 11: Test case 3: comparison of the solutions produced by Methods 1, 2, 3 in two norms: $L^2(I;H^1(\Omega))$ (left) and $H^1(I;H^{-1}(\Omega))$ (right).

Test case	1	2	3
Method 2	0.92	1.01	1.02
Method 3	1.17	0.82	0.38

Table 1: Relative CPU times for Methods 2 and 3 with respect to Method 1 for the three test cases.

a dominant non-selfadjoint part, as in advection-dominated transport problems.

References

[1] Andreev, R. *Stability of space-time Petrov–Galerkin discretizations for parabolic evolution equations.* PhD thesis, ETH Zürich, 2012.

[2] Andreev, R. Stability of sparse space-time finite element discretizations of linear parabolic evolution equations. *IMA J. Numer. Anal. 33*, 1 (2013), 242–260.

[3] Andreev, R. Space-time discretization of the heat equation. *Numer. Algorithms 67*, 4 (2014), 713–731.

[4] Barrault, M., Maday, Y., Nguyen, N. C., and Patera, A. T. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *C. R. Math. Acad. Sci. Paris 339*, 9 (2004), 667–672.

[5] Cancès, E., Ehrlacher, V., and Lelièvre, T. Convergence of a greedy algorithm for high-dimensional convex nonlinear problems. *Math. Models Methods Appl. Sci. 21* (2011), 2433–2467.

[6] Cancès, E., Ehrlacher, V., and Lelièvre, T. Greedy algorithms for high-dimensional eigenvalue problems. *Constructive Approximation 40*, 3 (2014), 387–423.

[7] Chinesta, F., Keunings, R., and Leygue, A. *The Proper Generalized Decomposition for Advanced Numerical Simulations.* Springer Briefs in Applied Sciences and Technology. Springer, Cham, 2014. A primer.

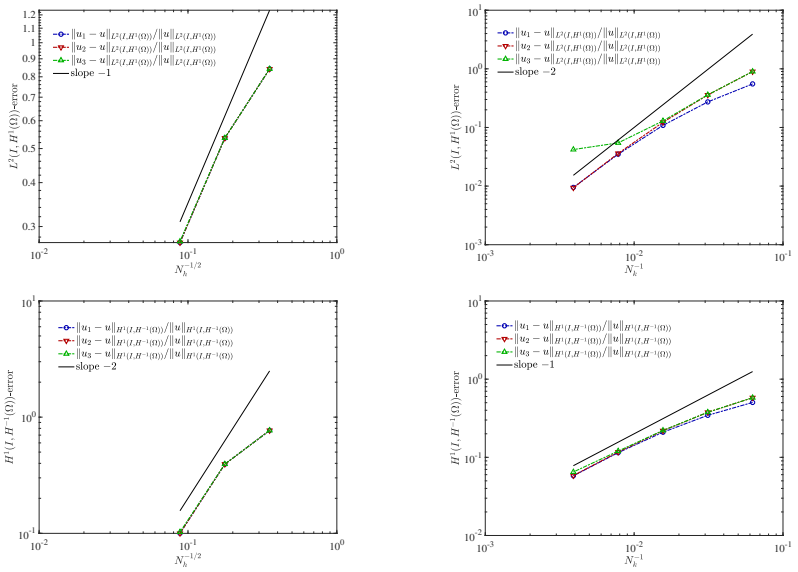

Figure 12: Test case 3: convergence study for Methods 1, 2 and 3 for errors measured in the $L^2(I;H^1(\Omega))$-norm (top row) and in the $H^1(I;H^{-1}(\Omega))$-norm (bottom row) for various mesh-sizes $N_h^{-1/2}$ (left column) and various time-steps N_k^{-1} (right column); in the right column, the curve for Method 2 overlaps with that of Method 1 for the smaller time-steps and with that of Method 3 for the larger time-steps.

[8] Dautray, R., and Lions, J.-L. *Mathematical Analysis and Numerical Methods for Science and Technology. Vol. 5. Evolution problems, I.* Springer-Verlag, Berlin, Germany, 1992.

[9] Ern, A., and Guermond, J.-L. *Theory and Practice of Finite Elements*, vol. 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[10] Ern, A., Smears, I., and Vohralík, M. Guaranteed, Locally Space-Time Efficient, and Polynomial-Degree Robust a Posteriori Error Estimates for High-Order Discretizations of Parabolic Problems. *SIAM J. Numer. Anal. 55*, 6 (2017), 2811–2834.

[11] Falcó, A., and Nouy, A. A proper generalized decomposition for the solution of elliptic problems in abstract form by using a functional Eckart-Young approach. *J. Math. Anal. Appl. 376*, 2 (2011), 469–480.

[12] Falcó, A., and Nouy, A. Proper generalized decomposition for nonlinear convex problems in tensor Banach spaces. *Numer. Math. 121*, 3 (2012), 503–530.

[13] Gander, M. J., and Vandewalle, S. Analysis of the parareal time-parallel time-integration method. *SIAM J. Sci. Comput. 29*, 2 (2007), 556–578.

[14] Gander, M. J., and Zhao, H. Overlapping Schwarz waveform relaxation for the heat equation in n dimensions. *BIT 42*, 4 (2002), 779–795.

[15] Giladi, E., and Keller, H. B. Space-time domain decomposition for parabolic problems. *Numer. Math. 93*, 2 (2002), 279–313.

[16] Griebel, M., and Oeltz, D. A sparse grid space-time discretization scheme for parabolic problems. *Computing 81*, 1 (2007), 1–34.

[17] Gunzburger, M. D., and Kunoth, A. Space-time adaptive wavelet methods for optimal control problems constrained by parabolic evolution equations. *SIAM J. Control Optim. 49*, 3 (2011), 1150–1170.

[18] Hoang, V. H., and Schwab, C. Sparse tensor Galerkin discretization of parametric and random parabolic PDEs—analytic regularity and generalized polynomial chaos approximation. *SIAM J. Math. Anal. 45*, 5 (2013), 3050–3083.

[19] Janssen, J., and Vandewalle, S. Multigrid waveform relaxation of spatial finite element meshes: the continuous-time case. *SIAM J. Numer. Anal. 33*, 2 (1996), 456–474.

[20] Kieri, E., Lubich, C., and Walach, H. Discretized dynamical low-rank approximation in the presence of small singular values. *SIAM J. Numer. Anal. 54*, 2 (2016), 1020–1038.

[21] Koch, O., and Lubich, C. Dynamical low-rank approximation. *SIAM J. Matrix Anal. Appl. 29*, 2 (2007), 434–454.

[22] Ladevèze, P. Nonlinear Computational Structural Mechanics: New Approaches and Non-Incremental Methods of Calculation, 2012.

[23] Le Bris, C., Lelièvre, T., and Maday, Y. Results and questions on a nonlinear approximation approach for solving high-dimensional partial differential equations. *Constr. Approx. 30*, 3 (2009), 621–651.

[24] Lions, J.-L., Maday, Y., and Turinici, G. Résolution d'EDP par un schéma en temps "pararéel". *C. R. Acad. Sci. Paris Sér. I Math. 332*, 7 (2001), 661–668.

[25] Lions, J.-L., and Magenes, E. *Non-homogeneous boundary value problems and applications. Vols. I, II.* Springer-Verlag, New York-Heidelberg, 1972. Translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181-182.

[26] Lubich, C., and Oseledets, I. V. A projector-splitting integrator for dynamical low-rank approximation. *BIT 54*, 1 (2014), 171–188.

[27] Mantzaflaris, A., Scholz, F., and Toulopoulos, I. Low-rank space-time isogeometric analysis for parabolic problems with varying coefficients. *Comp. Methods Appl. Math.* (2018). Published online, DOI `https://doi.org/10.1515/cmam-2018-0024`.

[28] Neumüller, M., and Smears, I. Time-parallel iterative solvers for parabolic evolution equations. arXiv:1802.08126, 2018.

[29] Nouy, A. Recent developments in spectral stochastic methods for the numerical solution of stochastic partial differential equations. *Arch. Comput. Methods Eng. 16*, 3 (2009), 251–285.

[30] Nouy, A. A priori model reduction through Proper Generalized Decomposition for solving time-dependent partial differential equations. *Comput. Methods Appl. Mech. Engrg. 199*, 23-24 (2010), 1603–1626.

[31] Paige, C. C., and Saunders, M. A. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software 8*, 1 (1982), 43–71.

[32] Schwab, C., and Stevenson, R. Space-time adaptive wavelet methods for parabolic evolution problems. *Math. Comp. 78*, 267 (2009), 1293–1318.

[33] Tantardini, F., and Veeser, A. The L^2-projection and quasi-optimality of Galerkin methods for parabolic equations. *SIAM J. Numer. Anal. 54*, 1 (2016), 317–340.

[34] Temlyakov, V. N. Greedy approximation. *Acta Numer. 17* (2008), 235–409.

[35] Thomée, V. *Galerkin finite element methods for parabolic problems*, second ed., vol. 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2006.

[36] Urban, K., and Patera, A. T. A new error bound for reduced basis approximation of parabolic partial differential equations. *C. R. Math. Acad. Sci. Paris 350* (2012).

[37] Uschmajew, A. Local convergence of the alternating least squares algorithm for canonical tensor approximation. *SIAM J. Matrix Anal. Appl. 33*, 2 (2012), 639–652.

[38] Wloka, J. *Partial differential equations*. Cambridge University Press, Cambridge, 1987. Translated from the German by C. B. Thomas and M. J. Thomas.